Nucleic Acids Research, Vol. 20, No. 6 1363–1370

Statistical analyses of counts and distributions of restriction sites in DNA sequences

Samuel Karlin, Chris Burge and Allan M.Campbell[1]
Department of Mathematics and [1]Department of Biology, Stanford University, Stanford, CA 94305, USA

Received October 28, 1991; Revised and Accepted January 22, 1992

ABSTRACT

Counts and spacings of all 4- and 6-bp palindromes in DNA sequences from a broad range of organisms were investigated. Both 4- and 6-bp average palindrome counts were significantly low in all bacteriophages except one, probably as a means of avoiding restriction enzyme cleavage. The exception, T4 of normal 4- and 6-palindrome counts, putatively derives protection from modification of cytosine to hydroxymethylcytosine plus glycosylation. The counts and distributions of 4-bp and of 6-bp restriction sites in bacterial species are variable. Bacterial cells with multiple restriction systems for 4-bp or 6-bp target specificities are low in aggregate 4- or 6-bp palindrome counts/kb, respectively, but bacterial cells lacking exact 4-cutter enzymes generally show normal or high counts of 4-bp palindromes when compared with random control sequences of comparable nucleotide frequencies. For example, *E. coli*, apparently without an exact 4-bp target restriction endonuclease (see text), contains normal aggregate 4-palindrome counts/kb, while *B. subtilis*, which abounds with 4-bp restriction systems, shows a significant under-representation of 4-palindrome counts. Both *E. coli* and *B. subtilis* have many 6-bp restriction enzymes and concomitantly diminished aggregate 6-palindrome counts/kb. Eukaryote, viral, and organelle sequences generally have aggregate 4- and 6-palindromic counts/kb in the normal range. Interpretations of these results are given in terms of restriction/methylation regimes, recombination and transcription processes, and possible structural and regulatory roles of 4- and 6-bp palindromes.

INTRODUCTION

The Kohara physical map (1) for *E. coli* was constructed using partial digestion with eight 6-cutters, 7 of which recognize exact 6-bp palindromic sites, *Bam*HI (GGATCC) — 470 occurrences, *Eco*RI (GAATTC) — 613, *Eco*RV (GATATC) — 1159, *Hin*dIII (AAGCTT) — 518, *Kpn*I (GGTACC) — 497, *Pst*I (CTGCAG) — 848, *Pvu*II (CAGCTG) — 1435, and an eighth, *Bgl*I ($GCCN_5GGC$) — 1572. Analysis of the counts and spacings of these restriction sites reveals two features: (i) There is substantial diversity in counts for the various enzyme sites, (ii) The spacings between sites of each type considered separately appear homogeneous, consistent with a uniform random distribution (2,3).

These observations prompted some more general questions: (1) What are the counts and spacings of all hexanucleotide palindromes (abbreviated 6-palindromes) in *E. coli* sequences? (2) Can the observed variation in the counts of 6-palindromes be explained on the basis of either mono, di, or trinucleotide frequencies? (3) How are 6-palindromes distributed in other DNA sequences? (4) Is there evidence of clumping (several short contiguous intervals between sites), overdispersion (long intervals) or excessive evenness (too few short and too few long intervals)?

We investigated counts and spacings of 4- and 6-palindromes across all the available *E. coli* DNA sequences (exceeding 1.43 Mb), as well as for sequences from several bacteriophages, bacteria, viruses, eukaryotes, and organelles (see Table 1). Restriction site counts and spacings in various species may suggest choices of restriction endonucleases to produce optimal fragment lengths in developing viable clone libraries, e.g., see (4).

MATERIALS AND METHODS

Data

Table 1 lists the data sets investigated. These include (accessible in EMBL) several large complete phage, viral and organelle genomes and eukaryotic sequences covering a broad phylogenetic range (mammalian, avian, amphibian, invertebrate, plant, and fungal representatives) and diverse eubacterial sequences including Gram-negative and Gram-positive species. The results with complete phage, viral, and organelle genomes are obviously free from bias. The *E. coli* conglomerate is an ensemble of disjoint contigs (5). The other phage, bacterial, *C. elegans* and *N. crassa* sequences were culled of redundant entries. The species collections were compiled from all available sequences in EMBL. Large unduplicated samples from the eukaryotic sequences were analyzed and had aggregate 4- and 6-palindrome counts/kb consistent with those of the total sets. The species sequence collections undoubtedly have biases. For example, the human sequences frequently center on genes of medical interest, the *Drosophila* set includes a plethora of genes acting in embryogenesis and development, and many of the *Rhizobium meliloti* sequences relate to nitrogen fixation.

Statistical controls

We use four types of controls in our assessments and interpretations of 6-palindromic heterogeneity: (i) distributional counts of the 64 possible 6-palindromes in each of 100 shuffled sequences generated by sampling without replacement from the parental sequence, (ii) distributions of the 64 palindromes in 100 random sequences, length 50 kb each, generated using the same nucleotide frequencies as the parental sequence, (iii) distributions of several sets of 64 randomly generated 6-words (hexanucleotides) in the parental sequence, and (iv) comparisons of counts of all 6-palindromes among the different data sets.

Table 1. Data

Organism	Length (kb)	C + G%[c]			
Bacteriophages					
T7[a]	39.9	48.40			
Lambda[a]	48.5	49.85			
PZA[a]	19.4	39.67			
P1 (33% of genome)	29.4	40.80			
T4 (50% of genome)	103.1	35.66			
Bacteria					
Gram− (α-purple group)					
Rhizobium meliloti	67.3	60.69			
Rhodobacter capsulatum	58.4	64.34			
Agrobacterium tumefaciens	46.9	51.63			
Gram− (β-purple group)					
Neisseria gonorrhoeae	68.9	51.34			
Gram− (γ-purple group)					
Escherichia coli	1,431.7	51.57			
Pseudomonas aeruginosa	104.9	63.15			
Haemophilus influenzae	32.8	37.17			
Gram+					
Bacillus subtilis	142.0	43.47			
Streptomyces lividans	20.0	68.74			
Thermus thermophilus	26.1	67.18			
Human Viruses[a]					
Adeno	35.9	55.20			
Cytomegalo (CMV)	229.4	57.16			
Epstein-Barr (EBV)	172.3	59.94			
Herpes Simplex 1 (HSV1)	152.3	68.73			
Varicella-Zoster (VZV)	124.9	46.02			
Vaccinia	191.7	33.40			
Eukaryotes					
Saccharomyces cerevisiae[b]	1,284.2	38.56			
Neurospora crassa	204.4	52.75			
Caenorhabditis elegans[b]	311.9	40.20			
Drosophila melanogaster[b]	1,432.7	45.73			
Xenopus laevis[b]	659.5	44.91			
Chicken[b]	1,001.4	50.27			
Human (20% of EMBL)	1,410.9	50.99			
Zea mays[b]	395.5	50.20			
Chloroplasts[a]					
Rice	134.5	38.99			
Tobacco	155.8	37.85			
Mitochondria[a], [c]		A%	C%	G%	T%
Paramecium aurelia	40.5	25.28	21.91	19.33	33.48
Paracentrotus lividus	15.7	30.77	22.51	17.18	29.54
Drosophila yakuba	16.0	39.49	12.17	9.25	39.10
Xenopus laevis	17.6	33.05	23.49	13.50	29.96
Rat	16.3	34.09	26.22	12.46	27.23
Human	16.6	30.92	31.24	13.13	24.71

[a]complete genome; [b]all of current EMBL; [c]in chromosomal, viral, and chloroplast DNAs, A ≅ T and C ≅ G in each strand. For mitochondrial DNAs, the differences are significant, so the composition of one strand is shown. See Methods concerning data selection and cleaning.

Expectations of counts of palindromes based on mono-, di-, and trinucleotide frequencies

For a model with the nucleotide occurrences independently distributed, the probability of observing a 6-word at any specified location is the product of the frequencies of the component letters. For a dinucleotide (Markov-immediate neighbor dependence) model, the probability of observing a particular 6-word, say w = ACCTAG, is $f_w = (f_{AC}\, f_{CC}\, f_{CT}\, f_{TA}\, f_{AG})/f_C\, f_C\, f_T\, f_A$. For a trinucleotide Markov model the probability is estimated by $f_w = (f_{ACC}\, f_{CCT}\, f_{CTA}\, f_{TAG})/f_{CC}\, f_{CT}\, f_{TA}$. In all cases the expected count would be $N\, f_w$ (designated C^{mono}, C^{di}, C^{tri}, depending whether the predictions are based on mono-, di- or trinucleotide frequencies, respectively) where N is the length of the sequence. These formulas have been widely used, (e.g., (6−8)).

Associations of palindromic counts

The standard measure of concordance is the cross (Pearson) correlation formula which can be confounded by outlier observations. The Kendall-Tau correlation coefficient is less affected in this way. Let N_1, (g), N_2 (g), . . . , N_{64} (g) be the respective counts of the 6-palindromes observed in genomic sequence g. Similarly, let $\{N_i(h)\}$ be the counts found in genomic sequence h. For each pair of 6-palindromes labeled i and j $(1 \le i < j \le 64)$ we determine

$$\tau_{ij} = \begin{cases} +1 & \text{if} \quad [N_i(g) - N_j(g)][N_i(h) - N_j(h)] > 0 \\ -1 & \text{if} \quad [N_i(g) - N_j(g)][N_i(h) - N_j(h)] < 0 \\ 0 & \text{if either} \quad N_i(g) = N_j(g) \text{ or } N_i(h) = N_j(h) \end{cases}$$

The Kendall-Tau association measure is

$$[1] \qquad \tau(g,h) = \frac{2\sum_{i \ne j} \tau_{ij}}{\left(\sqrt{n(n-1) - 2t}\sqrt{n(n-1) - 2s}\right)}$$

where t and s are the number of ties among pairs of N_i (g),$N_j(g)$ and N_i (h),N_j (h), respectively, and $n = 64$. Clearly $-1 \le \tau \le 1$ and $\tau = 1(-1)$ if and only if the palindromic counts in sequence g exhibit a completely concordant (discordant) ordering to the palindromic counts in sequence h. A value $|\tau| \ge 0.40$ for two random orderings (of 64 distinct real numbers) has a probability $< 10^{-2}$ of occurring.

Given expected counts of palindromes based on mono-, di-, or trinucleotide frequencies, the correlation statistics of the expectations with the observed counts are determined by the calculations of Eq. [1] using $\{C_i(g)\}$ versus $\{C_i{}^{\text{mono}}(g)\}$, $\{C_i{}^{\text{di}}(g)\}$ or $\{C_i{}^{\text{tri}}(g)\}$, respectively.

Extremal spacings of a marker

Consider a sequence of length K and a specified word type with k occurrences randomly distributed in the sequence. These induce $k + 1$ spacings, $(U_0, U_1, \ldots, U_k)$ where U_i is the distance (numbers of units) from the ith occurrence to the $i + 1$st occurrence, U_0 is the distance before the first occurrence, and U_k that after the last. Distances are scaled so that one unit equals $1/K$. Our statistical analysis focuses on the extremal spacings $m = \min\{U_0, U_1, \ldots, U_k\}$ and $M = \max\{U_0, U_i, \ldots, U_k\}$. The following classical distributional formulas (e.g., (9)) of independent uniformly distributed points on the unit interval serve in the analysis of the spacings of a marker:

$$[2] \qquad F(a) = \text{Prob}\{m \ge a\} = [1 - (k+1)a]^k \quad , \quad 0 < a \le \frac{1}{k+1}$$

[3] $$G(b) = \text{Prob}\{M \le b\} = \sum_{i=0}^{k+1} \binom{k+1}{i} (-1)^i [(1 - bi)_+]^k \quad \text{for} \quad b \ge \frac{1}{k+1}$$

where $(1 - bi)_+ = (1 - bi)$ if $bi < 1$ and 0 otherwise. The criterion for an extreme minimum at the 1% significance level involves the determination of a^* such that $F(a^*) = .99$, and for an observed m smaller than a^* the minimum spacing is considered significantly small. Similarly, the largest gap is significantly large if the observed M exceeds b^* where b^* satisfies $G(b^*) = .99$. For m too large and/or M too small the spacings are considered to be overly even. The formulas [2] and [3] apply to k sites on a linear string; if the sites are sampled equally likely on a circular string (genome), the formulas need to be adjusted by replacing k by $k - 1$.

RESULTS

Counts of 4- and 6-palindromes

Table 2a compares the average counts/kb of all 4- and 6-palindromes across a broad spectrum of genomic sequences. The range of corresponding average palindromic counts/kb for 100 random 50 kb sequences is recorded in Table 2b and typical

Table 2a. Average counts per kb of 4- and 6-base palindromes in all studied organisms and 20 shuffled sequences* for some organisms

Organism	mean # of 6-palindromes per kb‡	# of restriction systems with exact 6-palindromic specificities	mean # of 4-palindromes per kb‡	# of restriction systems with exact 4-palindromic specificities
Bacteriophages				
T7	6.6 (14.4–17.2)*		(41.9 (59.2–65.4)*	
Lambda	9.5 (14.3–17.0)*		54.9 (59.3–65.5)*	
P1	13.5		57.2	
PZA	11.6 (14.9–19.7)*		53.7 (63.1–70.8)*	
T4	16.5		63.9	
Bacteria				
R. meliloti	20.1	0	79.5	0
R. capsulatum	18.4	1	86.5	0
A. tumefaciens	16.8	2	67.7	0
N. gonorrhoeae	11.6	2	50.1	2
E. coli	10.1	17	59.6	0
P. aeruginosa	18.2	6	75.8	0
H. influenzae	14.0	1	56.8	4
B. subtilis	13.4	5	55.9	3
S. lividans	20.9	0	82.4	0
T. thermophilus	15.9	0	58.3	1
Large human viruses				
Adenovirus	14.4 (14.4–18.4)*		61.9 (60.9–67.5)*	
CMV	15.2 (15.1–17.2)*		65.1 (63.8–68.1)*	
EBV	15.1 (15.3–18.1)*		55.8 (66.0–68.9)*	
HSV1	19.6 (19.1–23.1)*		77.5 (77.3–81.3)*	
VZV	15.7 (15.0–16.9)*		66.5 (61.6–65.5)*	
Vaccinia	20.4 (20.2–21.9)*		68.9 (73.3–78.0)*	
Eukaryotic species				
S. cerevisiae	16.6		58.3	
N. crassa	13.9		53.0	
C. elegans	17.2		58.0	
D. melanogaster	18.3		64.8	
Xenopus laevis	15.7		54.3	
Chicken	14.0		50.6	
Human	13.7		48.6	
Maize	18.9		64.7	
Chloroplasts				
Rice	15.2 (17.1–18.8)*		58.1 (65.8–68.7)*	
Tobacco	15.7 (17.4–19.1)*		58.5 (67.8–69.8)*	
Mitochondria				
Paramecium aurelia	18.7 (15.1–17.8)*		61.2 (60.4–66.6)*	
Paracentrotus lividus	19.3 (16.1–20.4)*		52.6 (62.6–69.5)*	
Drosophila yakuba	36.5 (33.3–37.8)*		101.3 (99.1–107.2)*	
Xenopus laevis	19.8 (16.1–20.4)*		65.2 (63.5–69.5)*	
Rat	17.1 (14.2–18.3)*		56.7 (59.0–66.1)*	
Human	14.3 (12.3–14.9)*		47.6 (51.6–56.5)*	

*Shuffled sequences, compare also with Tables 1 and 2b.

‡Because many sequences (other than complete genomes) begin and/or end with restriction sites, these calculations were also performed excluding the first 6 bp and the last 6 bp of each sequence; counts of 4- and 6-palindromes were typically slightly lower but never by more than about 2%.

Table 2b. Counts per kb of 4- and 6-palindromes in 100 randomly generated sequences each of length 50 kb

Frequencies	6-palindromes Mean	Range	4-palindromes Mean	Range
A=T=C=G=25%	15.5	14.3–17.1	62.3	59.2–64.8
A=T=22.5%, C=G=27.5%[a]	16.1	14.5–18.4	63.8	60.9–67.1
A=T=20%, C=G=30%[a]	17.6	15.9–19.2	67.5	63.1–70.8
A=T=17.5%, G=C=32.5%[a]	20.3	18.7–21.9	74.3	71.4–77.0
A=T=15%, G=C=35%[a]	24.3	22.6–26.9	83.9	80.7–87.3
A=T=12.5%, G=C=37.5%[a]	30.4	28.5–32.5	97.5	93.5–100.7
A=T=10%, G=C=40%[a]	39.3	37.0–41.2	115.5	111.1–119.1

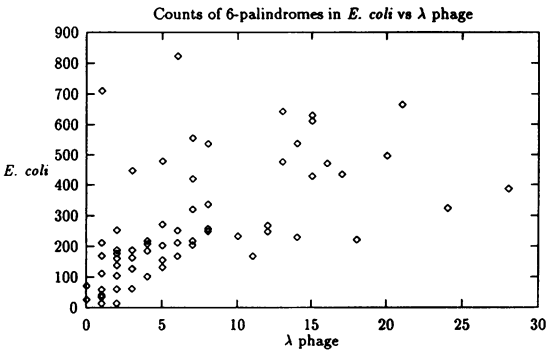
[a]or replace A by G and T by C.

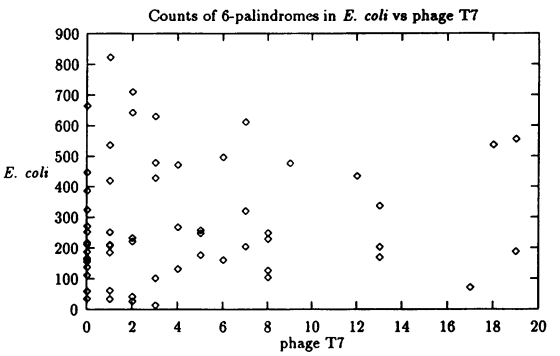
counts for shuffled sequences are also given in Table 2a. Average 6-palindrome counts are significantly low in all the bacteriophage sequences except T4 and drastically low in T7. Sharp (10) ascertained the numbers of all 6-palindromes and restriction sites in T7, λ, and in several small coliphage (φX174, G4, fl, fd, φ29, IKe) and noted pervasive low counts, interpreting this outcome as restriction avoidance.

The aggregate 4- and 6-palindrome counts are significantly low in the two chloroplast genomes examined and the same holds for 4-palindromes in the mammalian mitochondrial genomes. Average counts of 6-palindromes in most eukaryotic sequence sets and human viral genomes generally fall into the random range, although the average 4-palindrome counts tend to the low side in these sequences. The 4- and 6-palindrome counts for the bacterial sequences are varied and perplexing. In particular, the α-purple group (*R. capsulatum, R. meliloti, A. tumefaciens*) have, on average, 4-palindrome counts that are singularly high, whereas the aggregate 6-palindrome/kb counts in these species are in the random range. The 4-palindrome counts in *P. aeruginosa* are high normal. On the other hand, *N. gonorrhoeae, H. influenzae, T. thermophilus*, and *B. subtilis* show significantly low aggregate 4- and 6-palindrome counts/kb. Unlike the 6-palindromes, the average counts of 4-palindromes for *E. coli* sequences conform to random expectations, and the same is true of the Gram-positive *S. lividans* sequences. These disparate outcomes show no clear relationship with genomic G+C content. However, the extent of over, normal, and under-representation of the aggregate numbers of 4- and 6-palindromes/kb in relation to the bacterial inventory of restriction systems targeted to 4- and 6-bp specificities, respectively, suggest a consistent pattern (see Discussion).

Extremal representations

For the aggregate *E. coli* sequences about 60% of individual 6-palindromes entail counts far below the expected value of 351; consult Figure 1. Inordinately low are *Xba*I (TCTAGA — 13 occurrences), *Spe*I (ACTAGT — 25), *Nhe*I (GCTAGC — 32), and *Avr*II (CCTAGG — 13), which all center on the tetranucleotide CTAG, by far the rarest 4-word in *E. coli* with a frequency about 0.0002. The G + C 6-palindromes *Nae*I, GCCGGC (40 occurrences) and *Apa*-I GGGCCC (34) are the next lowest, which contrast sharply with the numbers of *Bss*HII GCGCGC-715 and CGCGCG-643, the most frequent 6-palindromes (only distinct nonoverlapping occurrences are counted once). The third most abundant 6-palindrome is the *Eco*RV-site (GATATC-660) and then *Pvu*II (CAGCTG-622). The latter two were used in the course of generating the Kohara





physical map. These extremely high and low individual palindrome occurrences are approximately similarly ranked in phage λ and several other bacterial sequences (data not shown).

Among the 4-palindromes, CTAG has the lowest or near lowest count in a broad spectrum of prokaryotic and eukaryotic sequences including λ, P1, *all* bacterial genomes examined, HSV1, CMV, VZV, *Drosophila*, chicken, and *C. clegans* and is unambiguously below average in most of the other sequences (Table 3). The 4-palindrome GATC (DAM methylase site) is extremely low in frequency in T7 (0.0002) compared to .0010 for CTAG, the second lowest. GGCC is under-represented in several bacteriophage genomes: T4, P1, PZA. For perspectives on the rarity of CTAG, GATC, and GGCC in enterobacterial sequences, see (11), and our Discussion.

Counts of long palindromes

The numbers, scaled to 50 kb, of exact palindromes of lengths 4, 6, 8, 10, 12, . . ., for each of the data sets were ascertained (data not shown, see also (12)). These were compared with the

Table 3. Extreme high and low frequency 4bp palindromes in phage, prokaryotic, viral, eukaryotic, and mitochondrial sequences

Number of organisms in each class in which given 4-palindrome is significantly rare (−) or significantly frequent (+)[a]					
4-Palindrome	Phage (5)	Prok.[c] (10)	Viral[c] (6)	Euk.[d] (8)	Mito.[e] (6)
AATT	1+	3−/1+	1−		3+
ACGT				2−	3−
AGCT					
ATAT		4−	1−/1+		1+
CATG		1−			
CCGG	4−	1−/3+	1−/1+	2−	6−
CGCG	2−	1−/3+	1−/1+	4−	6−
CTAG[b]	3−	9−	4−	3−	1−
GATC	2−	1−/1+			1−
GCGC	1−	1−/4+	1−/1+	3−	5−
GGCC	3−	1−/4+	1−/1+	1−	1−
GTAC		3−			1−
TATA		5−	2−/1+		1+
TCGA			2−	3−	1−
TGCA		4−			1−
TTAA	1+	1+	1−		3+

[a]Significantly low means scaled counts to 50 kb length ≤75.4 and significantly high means scaled counts to 50 kb length ≥504.2.
[b]CTAG consistently low, never high.
[c]The 1− and 1+ is *H. influenzae* (genome of G+C = 37%). The 1−/1+ in viruses corresponds to HSV1 (G+C = 68%) and vaccinia (G+C = 33.4%).
[d]No high 4-palindromes.
[e]Skewed genomes and short sequence lengths account for some of these numbers.

range of corresponding counts for 100 random sequences each of length 50 kb. The statistics reveal that the exact palindrome counts (for lengths from 6 to 14 bp) in T7, λ, and *E. coli* are significantly low. By contrast, counts in the eukaryotic sequences for these moderate length palindromes are generally in the range of the random samples. Long palindromic elements (head to tail ≥ 18 bp) abound in the herpesvirus genomes, often positioned proximal to viral origins of replication or in the embrace of promoter and enhancer elements.

Correlations of 6-palindromic counts between data sequences

For each sequence pair, the Kendall-Tau correlations were calculated (see Methods) relative to individual 6-palindrome counts. Significant 6-palindrome Kendall-Tau correlations are displayed in Table 4 only if at least one of the sequences involved has moderate G+C frequency (that is .45 ≤ G+C% ≤ .55), since otherwise the compositional biases dominate the results. For example, the sequences from P1, T4, PZA, *B. subtilis*, vaccinia, yeast, *Drosophila*, chloroplasts (rice and tobacco), and mitochondria (sea urchin, *Drosophila, Xenopus*, rat), all A+T rich (G+C% < .45), produce Kendall-Tau correlations ≥ .35 (mostly ≥ .4) with each other. Similarly, HSV1, EBV, CMV, adenovirus, and the G+C rich bacterial sequences yield high correlations concomitant to their high G+C biases. The two groups of high versus low G+C content, as expected, entail strong negative correlations. Apart from effects of compositional extremes, the following observations stand out: (i) 6-palindrome counts of the T7 genome do not correlate significantly with any other sequence. (ii) The 6-palindrome correlations of *E. coli* with the temperate phages λ (.65) and P1 (.36) are high, whereas correlations with the lytic phages T7 and T4 are not significant (in the range −.2 to .2). (iii) The human and chicken correlation

Table 4. Significant Kendall-Tau all 4-word and 6-palindrome count correlations (1 † ≥ .35) for sequence pairs with at least one sequence having compositional G+C% between .45 and .55[a]

Organisms[b]	all 4-words	6-palindromes
Lambda/*E. coli*	0.65	0.55
Lambda/*B. subtilis*	0.48	0.45
Lambda/P1		0.42
Lambda/*Drosophila*		0.45
P1/*E. coli*	0.40	0.36
P1/*Drosophila*	0.48	0.41
B. subtilis/*E. coli*		0.41
B. subtilis/*Drosophila*	0.50	0.53
Adeno/*Neurospora*	0.41	
Adeno/chicken	0.42	
Adeno/human	0.41	
EBV/*Neurospora*	0.43	
EBV/chicken	0.43	
EBV/human	0.49	0.43
Yeast/*Drosophila*	0.43	0.51
Human/chicken	0.85	0.78

[a]All pairs of sequences of mutually high G+C composition or mutually low G+C composition tend to have high correlations (≥ 0.35) and of high versus low composition correlation < (−0.35).
[b]Correlation of pairs not recorded and not of the category (a) are not statistically significant. Thus, T7 (G+C% = 48.4) has all correlation values with all other sequences between −0.3 and 0.3 (and mostly between −0.2 to 0.2).

is markedly high at 0.78. Results obtained for correlations evaluated with respect to all (256) 4-word counts are consistent with the 6-palindrome correlations (Table 4).

Correlations of 6-palindrome counts with expectations based on mono-, di-, or trinucleotide frequencies

A Markov predicted frequency can be calculated for a given word based on the observed mono-, di-, trinucleotide (or even higher order) frequencies in a sequence, as described in the Methods section. The Kendall-Tau correlations were ascertained for these Markov expected numbers in relation to the observed palindrome counts. Markov Order 2 (trinucleotide) predictions correlate moderately with counts and orderings of palindromes for *E. coli* and λ sequences but weakly for T7. Markov Order 0 (mono) and 1 (di) predictions are correlated not at all or weakly with respect to both the counts and orderings of 6-palindromes in these organisms (data not shown, compare to (3)).

Spacings of palindromes in the λ genome

The distribution of each 6-palindrome around the λ-genome was tested for clumping, overdispersion or unusual regularity (see Methods). As a further control, we assessed the spacings of 64 random 6-words. Of the individual 6-palindromes, four involved significantly long gaps (overdispersion): CAGCTG (15 copies) maximum gap M = 21299 bp, CATATG (7) M = 36001 bp, CCTAGG (2) M = 48428 bp, CTGCAG (28) M = 14057 bp, compared to two extremes for the random words. Clumping was revealed for a single palindrome (CCTAGG (2), M = 74 bp) compared to none for the random words.

Extremal tests on spacings of 6-palindromes in T7 did not reveal a single 6-palindrome with abnormal spacings. Also, the spacings of GATC in T7 are not unusual in any way. We further investigated the spacings of the 6-palindromes in the 72 kb human β-globin region. No significant clumping was encountered. A single long gap was observed for the palindrome AAATTT (56

occurrences) with the maximal fragment distance M = 14149 bp; the palindrome TAGCTA (10 occurrences) showed overly even spacings in the β-globin sequence.

The extreme rarity of CTAG in λ-phage and *E. coli* and of GATC in phage T7 prompted us to investigate more closely the distribution of these tetranucleotides in these three organisms. The occurrences of CTAG in the λ genome are distributed at the locations indicated (each digit covers 1 kb).

```
0         10        20        30        40        kb
00000000000000000000001411000000011010000001102O #CTAG/kb
```

Note that CTAG is missing in the left half of the λ genome in a segment of 24743 bp and occurrences concentrate in three clusters in the right half. The formulae for extremal spacings confirm that the distribution of CTAG sites in λ is nonrandom. Of the 14 occurrences of CTAG in the λ genome, 8 are located in noncoding regions or at stop codons, 4 in ORFs of undetermined expression, 1 in the CI gene near the carboxyl end, and 1 in gene S (affecting cell lysis).

Although it was difficult to interpret the extremal spacings formulae for the large collection of contigs which make up the *E. coli* data set (there is a natural bias toward small spacings due to many short contig sequences), we did observe one prevalent pattern: CTAG occurs much more frequently in rRNA genes than elsewhere. In each of the 7 *E. coli* rRNA genes there is a distinct cluster of 7−9 CTAG's over a length of approximately 4400 bp, or once about every 400−600 bp, whereas the mean frequency of CTAG in *E. coli* is only about one per 4700 bp. Is it possible that CTAG sites are nucleation or anchor points in the assembly of the ribosomal complex or that ancestral CTAGs are better conserved here because rDNA changes more slowly in evolution? We have already emphasized the prevalent low frequency of CTAG in bacterial genomes. However, relative to its low frequency, clustering of CTAG is apparent in the 16S and 23S ribosomal units of many bacteria: 16S of *A. tumefaciens* (length 1489 — 3 copies of CTAG), *P. aeruginosa* (1537 — 7), *S. lividans* (1531 — 3), *T. thermophilus* (2331 — 6), *B. stolpii* (1553 — 6); 23S of *R. capsulatum* (2884 — 5), *T. thermophilus* (2915 — 8) and *B. subtilis* (rrnB:23S, 26S, 5S: 7430 — 15). The oriC region of *B. subtilis* also contains a relative excess of CTAG. In most of these species the relative positions in the rRNA segments of the CTAG tetranucleotides are closely conserved.

DISCUSSION

Our principal findings can be summarized as follows:

(i) The aggregate counts of 6-palindromes are significantly low in all bacteriophages studied except T4. Counts of 6-palindromes in T7 are drastically skewed to the low end. Exact aggregate palindrome counts of all moderate lengths (6−14 bp) are significantly low in λ, T7, and *E. coli*. By contrast, numbers of close dyad pairings (stem lengths ≥ 5 bp, loop lengths between 5 and 30) are *not* under-represented in these sequences (12).

(ii) The average counts of 4- and 6-palindromes in bacterial sequences are strikingly disparate. In fact, unlike the diminished cumulative numbers of 6-palindromes, the average count of 4-palindromes in *E. coli* and in *S. lividans* sequences are not distinguishable from random expectations. *B. subtilis*, *N. gonorrhoeae*, *H. influenzae*, and *T. thermophilus* sequences are significantly low in aggregate 4- and 6-palindrome counts/kb, whereas *A. tumefaciens*, *R. capsulatum*, and *R. meliloti* sequences are significantly high in 4-palindrome counts/kb. These latter species sequences are confined to the normal range for 6-palindrome counts.

(iii) Although most individual 6-palindromes in the T7, λ, and *E. coli* sequences (more than 60%) have low counts, see Figure 1, there are a few 6-palindromes present at relatively high counts. The most prominent pattern is the very low occurrence of 6-palindromes that center on the extremely rare tetranucleotide CTAG. The results are similar with *Salmonella typhimurium*. For example, CCTAGG is extremely rare in *S. typhimurium* (only 10 copies) as verified by restriction digest (13). The two most frequent 6-palindromes in *E. coli* are the iterated GCGCGC and CGCGCG sites (also in the other bacterial sequences with random 6-palindrome counts). By contrast, GGGCCC and GCCGGC are very rare in *E. coli*, and $(AT)_3$ and $(TA)_3$ have average counts.

(iv) In most nuclear eukaryotic and viral sequences aggregate counts of 6-palindromes occur as if sampled from random sequences, and this is true of longer palindromes as well. However, the average numbers of 4-palindromes often tend to be under-represented. Also, the corresponding counts of 4- and 6-palindromes in chloroplast DNA are significantly low.

The foregoing observations prompt us to consider general mechanisms whereby palindromes as a group are selected against and compensating specific mechanisms that either spare certain palindromes or actively select for them.

Palindromes can isomerize to form cruciform structures, and recombination enzymes that nick such structures are known (14). The failure of many workers to clone large exact (≥ 30 bp) palindromes in *E. coli* (15, 16) has been traced to the presence of a single *E. coli* gene: *sbcC* (17). Apropos, it might be inferred from ref. (16) that human chromosomes can tolerate large palindromes which cannot be cloned into normal *E. coli*. If the *sbcC* product occasionally acted on palindromes as short as 6 to 14 bp, these palindromes would confer selective disadvantage. Energetically, 4-palindrome loopouts are prohibitive.

Most 6-palindromes are type II restriction sites (to date 53 of the 64 possible 6-palindromes and 13 of the 16 4-palindromes have been identified as restriction sites (18)). A type II restriction system entails a DNA methylase and an endonuclease targeted to the same specific sequence (4 to 8 bp length). Restriction systems in bacterial species primarily act to limit infection by bacteriophage. In opposition, many phage select out vulnerable restriction sites and develop versatile anti-restriction functions (e.g., SAMase in T3, *ral* gene of λ). If restriction mechanisms are selected for their ability to exclude phage, then the selective pressure on phage genomes to mutate away restriction sites may be substantial. The paucity of certain restriction sites in phage DNAs has been noted and discussed in these terms previously (e.g., (19, 20, 10)). The fact that 6-palindromes have normal occurrence in T4 accords with this explanation, because T4 DNA contains hydroxymethylcytosine (frequently glycosylated) rather than cytosine (21) and its DNA is therefore resistant to most restriction enzymes. The extreme low counts of 4- and 6-palindromes in T7 may suggest a broad range of historical hosts. In this vein, there are documented cases of phage which infect disparate bacteria (e.g., Mu-1 in *E. coli* and *Citrobacter freundii*). Bacterial DNA may also experience some selection against restriction sites, either because of occasional failure of methylation modification or of natural recombination resulting from intrastrain gene transfer.

Comparisons of diverse total counts of 4- and 6-palindromes in bacterial species are tantalizing. However, a consistent pattern emerges relating the degree of over, normal, and under-representation of the aggregate 4- and 6-palindrome counts/kb to the bacterial repertoire of 4- and 6-bp restriction system specificities, respectively. Table 2a includes a count of *distinct* exact 4- and 6-palindrome sites recognized by known restriction enzymes in the 10 bacterial species. In our further discussion we assume that the presently compiled restriction enzyme data base (of $\geq$700 entries (18)) reasonably reflects the nature and scope of the repertoire of restriction enzyme specificities available to each bacterial species analyzed in Table 2. It appears that under-, normal-, or over-representations of the aggregate count/kb of 4- and 6-palindromes in the various bacterial sequences correlate with their numbers of restriction systems as follows: Table 2 indicates that those bacteria without 4-bp target restriction enzymes possess a normal or an excess in 4-palindrome counts/kb. The same relationship applies to the δ-purple bacterium, *Myxococcus xanthus* (30 kb genomic sequence available) bearing no 4-bp restriction system and a significant excess of average 4-palindrome counts (data not shown). On the other hand, the bacterial species protected by multiple 4- and 6-cutter restriction systems are consistently significantly low in aggregate 4- and 6-palindrome counts/kb. This outcome prevails independently of genomic compositional biases as attested to by *B. subtilis, N. gonorrhoeae, H. influenzae*, and *T. thermophilus* (compare Tables 1, 2a and 2b).

In a survey of 172 *E. coli* isolates, no restriction endonucleases with an exact 4-bp target were characterized (18) concomitant with a normal 4-palindrome count/kb (Table 2). By contrast, *B. subtilis* sequences contain relatively many 4-bp restriction systems and a significantly diminished count of 4-palindromes. In view of the voluminous investigations of the *E. coli* genome, the nonexistence or paucity of exact 4-bp restriction enzymes in *E. coli* cells seems reliable. (The regulatory DAM methylation sequence GATC in *E. coli* is not a restriction site.) However, *E. coli* cells do employ many distinct 6-bp cutters (currently 17 identified) and have correspondingly low average 6-palindrome counts. From the foregoing facts and perspectives, we propose the hypothesis that the possession of one or more restriction enzymes of a given target length tends to result in reduced average counts of palindromes of that size, in part because of occasional failure of methylation modification as suggested above. In the absence of such selection, a normal average palindromic count would be expected, in agreement with most of our data. Why the aggregate 4-palindrome counts are inordinately high in the α-purple group, *R. capsulatum, A. tumefaciens, R. meliloti*, and the δ-purple *M. xanthus* remains a conundrum.

Exact palindromes and other very close dyads (potential loop length 0−4) may impede transcription (by creating polymerase pause sites) or correspondingly induce ribosome saltation events during translation. In *S. typhimurium* bacterial sequences there are documented transcriptional pause sites (22). The causes and mechanisms are unresolved. Some contend that RNA secondary structures (23) are the primary agent and others endorse RNA polymerase-DNA sequence interactions as decisive (24). Aspects of translational pause sites are discussed in (25, 26). Systematic studies of the effect of small palindromes on rates of transcription and/or translation elongation are not available. In both prokaryotes and eukaryotes, close dyads and palindromes are rarer in coding sequences than in noncoding sequences (27). Close dyads frequently occur in regulatory sites and flanking sequences, but are avoided in coding sequences where they might provide miscues. Palindromes can participate as recognition or binding sites for regulatory factors. In this context the under-representation might reduce errors in controls. Close dyads may also stabilize RNA transcripts against exonuclease degradation. In contrast to protein coding genes, tRNA and rRNA genes are rich in global and local secondary structure. Thus, there may be selection against both palindromes and close dyads within coding regions and selection for close dyads (potential stem length 4−30) but not for exact palindromes or very close dyads in noncoding regions.

(v) Many sequence attributes of *E. coli* and λ correlate strongly (Table 4), consistent with the thesis that they have coevolved for an extended period. Similarities between the genomes of temperate phages λ and P1 and their host *E. coli* DNA are seen in the high correlation coefficient for tetranucleotide counts and for total counts of 6-palindromes (Table 4). By contrast, corresponding correlations with the lytic phages T7 and T4 are not significant. Thus, the DNA of temperate phages may be subject to similar selection pressures as that of their host.

(vi) The frequency of CTAG appears very low in all bacterial sequences (Table 3) and substantially low in many eukaryotes and their viruses, including *Drosophila*, chicken, *C. elegans*, CMV, HSV1, and adenovirus (see Table 3). No convincing explanation is available. The perfect 14 bp palindrome ACTAGTTAACTAGT is the consensus binding site for the *trpR*-encoded repressor, and this important regulatory activity might require sufficient rarity of CTAG. Moreover, the stop codons TAG embedded in CTAG in opposite orientations may be selected against. However, the stop codon palindromic TTAA has normal representations in most organisms. The almost universal rarity of CTAG may implicate a structural role or defect. In this context, the crystallographic resolution of the *trpR*-DNA complex suggests that CTAG 'kinks' (28) which may, under conditions of supercoiling, be structurally deleterious. The potential role of the *vsr* gene product/VSP repair system (29) in reducing the frequency of CTAG and certain other DNA tetramers in certain bacterial genomes is discussed in (30, 31).

(vii) Although the frequency of the DAM methylation site G-ATC is often low in enterobacterial phage genomes (rare in T7, see ref. (32)), it is not significantly reduced in *E. coli*. In *E. coli* the role of the DAM methylase in repair, recombination, and replication may contribute to selection for sufficient GATC representation. This may also apply to the temperate phages, which have GATC at higher frequency than T7 or T4 (data not shown).

(viii) The scope of methylase activity in prokaryotic organisms is abundant and versatile. There are over 130 characterized DNA methyltransferases and over 240 restriction endonucleases with determined sensitivities to site-specific DNA modifications (33). Apart from inhibiting DNA cleavage by a restriction endonuclease, DNA methylation can interfere with many sequence specific DNA binding proteins and cause rate effects in restriction reactions and affect transcription and translation. Among 4-palindromes the following are established methylation sites in at least one bacterium or phage (33): AGCT, CATG, CCGG, CGCG, GATC, GCGC, GGCC, GTAC, TCGA but none act as restriction sites in *E. coli*. Under-representation of 4- and 6-palindromes and of certain other short oligonucleotides (30) is in part a concomitant of the hypermutable 5-methyl cytosine modification.

An analysis of restriction site data (18) reveals a pronounced

bias toward G+C rich target specificities. In fact, among exact 4-cutters cumulated over bacteria, the average G+C content of the sites was 3.2 out of 4 bases; average G+C content of exact 6-cutter sites was 4.6 out of 6 bases. What can account for this preference of G+C rich restriction endonuclease specificities? Possibly, the stronger hydrogen bonding associated with G:C base pairs would tend to produce a relatively more stable shape (contrasted to a looser shape of an A+T rich oligonucleotide), more easily recognized by a restriction enzyme and/or by the associated methylase and thereby providing better protection from invading phage. In this perspective, bacteria would evolve more restriction systems targeted to G+C rich oligonucleotides. Phage could reduce their vulnerability by evolving a more A+T rich genome. Interestingly, a survey of all coliphage sequences revealed no genome of G+C content exceeding 52%.

ACKNOWLEDGEMENTS

We are happy to acknowledge discussions and comments on the manuscript by Drs B.E.Blaisdell, D.Botstein, V.Brendel, M.McClelland, N.Murray, and C.Yanofsky. Supported in part by NIH Grants GMHG00335-03, GM10452-28, AI08573, and NSF Grant DMS86-06244.

REFERENCES

1. Kohara,Y., Akiyama,K. and Isoro,K. (1987) *Cell* **50**, 495-508.
2. Churchill,G.A., Daniels,D.L. and Waterman,M.S. (1990) *Nucleic Acids Res.* **18**, 589-597.
3. Karlin,S. and Macken,C. (1991) *Nucleic Acids Res.* **19**, 4241-4246.
4. Patel,Y., Van Cott,E., Wilson,G.G. and McClelland,M. (1990) *Nucleic Acids Res.* **18**, 1603-1607.
5. Rudd,K.E., Miller,W., Werner,C., Ostell,J., Tolstoshev,C. and Satterfield,S.G. (1991) *Nucleic Acids Res.* **19**, 637-647.
6. Blaisdell,B.E. (1985) *J. Mol. Evol.* **21**, 278-288.
7. Phillips,G.J., Arnold,J. and Ivarie,R. (1987) *Nucleic Acids Res.* **15**, 2611-2626.
8. Stückle,E.E., Emmrich,C., Grab,U. and Nielson,P.J. (1990) *Nucleic Acids Res.* **18**, 6641-6647.
9. Karlin,S. and Taylor,H. (1981) *A Second Course in Stochastic Processes*, Chap. 13. Academic Press, New York.
10. Sharp,P.M. (1986) *Mol. Biol. Evol.* **3**, 75-83.
11. McClelland,M., Jones,R., Patel,Y. and Neilson,M. (1987) *Nucleic Acids Res.* **15**, 5985-6008.
12. Karlin,S. (1986) In Karlin,S. and Nevo,E. (eds) *Evolutionary Processes and Theory*. Academic Press, FL, 329-363.
13. Wong,K. and McClelland,M. (1992) *J. of Bacteriol.* in press.
14. Mizuuchi,K., Kemper,B., Hays,J. and Weisberg,R. (1982) *Cell* **29**, 357-365.
15. Leach,D.R.F. and Stahl,F. (1983) *Nature* **305**, 448-451.
16. Wyman,A.R., Wolfe,L.B. and Botstein,D. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 2880-2884.
17. Chalker,A.F., Leach,D.R.F. and Lloyd,R.G. (1988) *Gene* **71**, 201-205.
18. Roberts,R. (1991) Restriction Enzyme Data Base. *Nucleic Acids Res.* April 1981.
19. Gunthert,U. and Trautner,T.A. (1984) In Trautner,T.A. (ed.), *Current Topics in Microbiology and Immunology*. Springer, Berlin, **108x**, 11-22.
20. Levin,B.R. (1986) In Karlin,S. and Nevo,E. (eds), *Evolutionary Processes and Theory*. Academic Press, New York, pp. 669-689.
21. Kornberg,A. (1980) *DNA Replication*. W.H.Freeman.
22. Levin,J.R. and Chamberlin,M.J. (1987) *J. Mol. Biol.* **196**, 61-84.
23. Landick,R. and Yanofsky,C. (1984) *J. Biol. Chem.* **259**, 11550-11557.
24. Arndt,K. and Chamberlin,M. (1990) *J. Mol. Evol.* **213**, 79-108.
25. Andersson,S.G.R. and Kurland,C.G. (1990) *Microbiol. Rev.* 198-210.
26. Randall,L.L., Josefsson,L.G. and Hardy,S.J.S. (1980) *Eur. J. Biochem.* **107**, 375-379.
27. Karlin,S., Morris,M., Ghandour,G. and Leung,M.-Y. (1988) *CABIOS* **4**, 41-51.
28. Otwinowski,Z., Schevitz,R.W., Zhang,R.G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) *Nature* **335**, 321-326.
29. Hennecke,F., Kolmar,H., Brundl,K. and Fritz,H.-J. (1991) *Nature* **353**, 776-778.
30. Burge,C., Campbell,A. and Karlin,S. (1992) *Proc. Natl. Acad. Sci. USA*, in press.
31. McClelland,M. and Bhagwat,A.S. (1992) *Nature* scientific correspondence, in press.
32. McClelland,M. (1985) *J. Mol. Evol.* **21**, 317-322.
33. Nelson,M. and McClelland,M. (1991) *Nucleic Acids Res.* **19**, 2045-2071.